Week 1 Homework (Due on Sept 24, 2017, 4:59PM)

1. Briefly explain what predictive modeling is, and why insurance industry needs it.
2. Explain the difference between the following pair of terms with at least one example for each type.
    a. Supervised model and unsupervised model
    b. Regression and classification
    c. Parametric model and non-parametric model
3. Get a list of all the functions in R that contains the string "***test"***
4. Write R command and calculate the following, including both R commands and results
    a. e^e
    b. $(2.3)^8 + ln(7.5) - \cos(\pi/2)$
    c. $|2^3-3^2|$

5. Generate the following sequences in R by using of command "seq" or "rep":
    a. 1 1 1 2 2 2 3 3 3 4 4 4 (quarterly dummy variable)
    b. 10.00000 10.04545 10.09091 10.13636 10.18182 10.22727 10.27273 10.31818 10.36364 10.40909 10.45455 10.50000 (number)
    c. "1" "apple" "2" "banana" "1" "apple" "2" "banana" (text)
6. R has a built-in data frame called "mtcars". Use commands to find the following info, including both R commands and results.
    a. Number of records
    b. Number of variables
    c. List the column "wt"
    d. Find the record call "Honda Civic"
    e. Find all records with mpg>30
    f. Find summary of the data frame
7. The "Third_party_claims" is an auto claim data. Please see the attach text file for more background information. The data file is a CSV file "Third_party_claims.csv". Your target variable is death rate.
   You may need to normalize the variables such that they will not be affect by population size. For example:
   tpc$mortality <- tpc$ki/tpc$population
   tpc$accidentRate <- tpc$accidents/tpc$population
   tpc$claimRate <- tpc$claims/tpc$population
   where tpc is your data frame. Please note that "sd", the statistical division, should be treated as a categorical variable.

    a. Build a linear model to calculate death rate by including only accidentRate, pop_density, and factor(sd).
    b. Justify the list of predictor variables by testing one or 2 additional variables.
    c. Plot residual vs. fitted values.
    d. Explain any issue you may find for the model and draw conclusion if assumptions for OSL are still valid.

   Please include R scripts, model results and necessary explanation.